T-61.3050 MACHINE LEARNING: BASIC PRINCIPLES, EXAMINATION

28 March 2009 at 10–13.

You must have passed the term project 2007 or part 1 of the term project 2008 to participate to this examination.

To pass the course you must also pass the term project. Results of this examination are valid for one year after the examination date.

To get full points you must choose and complete **five of the six problems**. Only the first five answers read by the examiner will be graded.

This examination has six problems (of which you must choose five) and three pages. You can answer in Finnish, Swedish or English. Please write clearly and leave a wide left or right margin. You can have a calculator, with memory erased. No other extra material is allowed.

An important grading criterion is understandability: in addition to being complete and correct, your answer should be understandable to your fellow student who has the necessary prerequisite knowledge but has not yet taken the course.

The results will be announced in Noppa on 28 April 2009, at latest. No other announcements will be sent.

You can keep this paper.

1. *Model selection.* Consider a process $h(x) = 2x^2 - x + 1$, where $x$ is a real-valued input variable. You have at your disposal only a data set of noisy samples from this process $\mathcal{X} = \{(r^t, x^t)\}_{i=1}^N$, where the noise is independent and identically distributed zero-mean Gaussian. The process is unknown to you, and the task is to estimate it.

   (a) Explain concepts "inductive bias", "underfitting", "overfitting", "hypothesis space" and "generalization", and their relation in the framework of this problem.

   (b) Give examples of hypothesis spaces that overfit and underfit the correct model.

   (c) How could you estimate the prediction error for yet unseen data?

2. *Principal Component Analysis.* Assume that your data $\mathcal{X}$ is $N$ d-dimensional real vectors, that is, $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^N$, $\mathbf{x}^t \in \mathbb{R}^d$. Consider the problem of reducing the dimensionality of your data to $k$ dimensions, where $k < d$, using principal component analysis (PCA).

   (a) What can PCA be used for? List some uses.

   (b) Write down in pseudocode how you could find the PCA representation of the data in $k$ dimensions. (Hint: it is probably easiest to use matrix representation here. You can assume that you have access to a function that gives eigenvectors and eigenvalues of a matrix.)

   (c) Give an interpretation of the PCA result in Figure 1. What can you tell about the original data points?

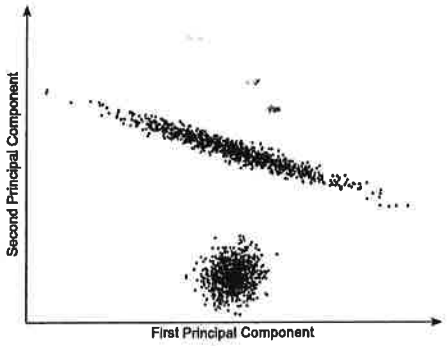


Figure 1: Principal Component Analysis result for problem 2.

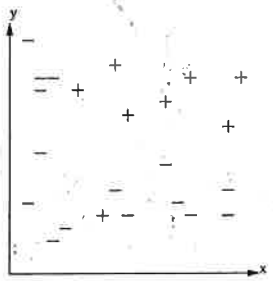


Figure 2: Toy data set for problem 5.

3. *Clustering.* Consider the problem of clustering $N$ real valued data vectors into $k$ clusters using the Lloyd's algorithm, also known as the $k$-means algorithm.

   (a) Write down the Lloyd's algorithm in pseudocode.

   (b) What can you say about the convergence and solutions found by the Lloyd's algorithm? How could you take this into account in practical data analysis?

4. *Nonparametric methods.*

   (a) What is the difference between parametric and nonparametric methods?

   (b) Describe the k-nearest neighbor (kNN) estimator in pseudo code.

   (c) What is the inductive bias in kNN?

5. *Decision trees.*

   (a) What is a decision tree? Define it.

   (b) Describe the ID3 algorithm by using pseudocode. Explain pruning in this context. Why and when is the pruning necessary?

(c) Sketch the running of the ID3 algorithm with a toy data set of Figure 2 (binary classification task in $\mathbb{R}^2$).

6. *Cross-validation.*

(a) What is cross-validation? Where and why is it used?

(b) Describe the k-fold cross-validation in pseudocode. (You can assume that you have access to functions that train and test a given model.)

(c) How to choose $k$? What is the valid range of values for $k$? What are the benefits and drawbacks of selecting $k$ to be the either end of the valid range?